Exploratory Data Analysis

Prof. Eric A. Suess

September 11, 2019

EDA

Today we will discuss Exploratory Data Analysis (EDA).

This is the process of exploring your data using visualization and transformations and modeling (will discuss modeling more later).

The goal of EDA is to develop a good understanding of your data.

Variables

- What variables are in the data set?
- What kind of variables are in the data set?
- What are the centers and spread of the variables?
- What are the relationships between the variables in the dataset?
- Is the data in a *tidy* format? Is it ready for further analysis and modeling?

Variables

Variables can be thought of in terms of two main types.

- Categorical (Ordinal or Nominal)
- Numeric (Continuous or Discrete)

Distributions

- One variable can be examined at a time.
- *Box-plots **or** histograms** are used to look at the distribution of numeric data.
- Are there *outliers* in the data?
- Are there any *misssing values* in the data?

Two variables

- Are there any relationships between pairs of variables in the data?